The Loflòc: A Morphological Lexicon for Occitan using Universal Dependencies

Marianne Vergez-Couret, Myriam Bras, Aleksandra Miletić, Clamença Poujade

1. Introduction and Objectives

Loflòc (Lexic obèrt flechit Occitan)

- Morphological lexicon for Occitan
- **680,000 entries** for **57,000 lemmas**
- Entries: **inflected form**, **lemma**, **part-of-speech tag** based on the **Universal Dependencies** framework
- **Lengadocian** dialect, **classical** spelling norm
- **Part of a wider drive** to **provide linguistic resources for Occitan**, which was a **low-resource** language at the outset of this endeavour
- **Overarching goals**: **preservation and dissemination of linguistic heritage** and **creation of resources suitable for the development of NLP tools**, in particular for basic processing such as lemmatization and morphosyntactic and syntactic analysis

2. Beginnings of NLP for Occitan

- **Occitan**: **Romance language** spoken in a large area in the south of **France**, in several valleys in **Italy** and in the Aran valley in **Spain**

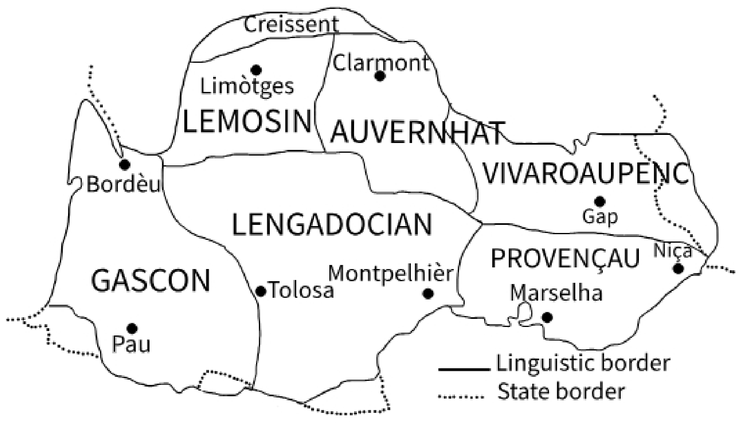

Figure 1: Occitan dialect continuum

- **Not standardized** as a whole
- **Six recognised varieties** or dialects which form a continuum: Auvernhàs, Gascon, Lengadocian, Lemosin, Provençau and Vivaro-Aupenc
- NLP challenges : **dialectal and spelling variation**

Nòrma classica (transcripcion)	**Nòrma mistralenca (originau)**
Mirèlha, Cant I (F. Mistral)	*Mirèio, Cant I (F. Mistral)*
Cante una chata de Provença.	Cante uno chato de Prouvènço.
Dins leis amors de sa jovença,	Dins lis amour de sa jouvènço,
A travèrs de la Crau, vèrs la mar, dins lei blats,	A travès de la Crau, vers la mar, dins li blad,
Umble [Umil] escolan dau grand *Omèra* [Omèr],	Umble escoulan dóu grand Oumèro,
Ieu la vòle seguir. Coma èra	Iéu la vole segui. Coume èro
Ren qu'una chata de la tèrra,	Rèn qu'uno chato de la terro,
En fòra de la Crau se n'es gaire parlat.	En foro de la Crau se n'es gaire parla.

Figure 2: Provençau dialect, two spelling norms
(source: `https://oc.wikipedia.org/wiki/Proven%C3%A7au`)

3. Loflòc

POS	meaning	count
ADJ	adjective	42,657
ADP	adposition	111
ADP+DET	adp.+determiner	22
ADV	adverb	1,170
AUX	auxiliary verb	184
CCONJ	coord. conjunction	9
DET	determiner	125
INTJ	interjection	189
NOUN	common noun	66,095
NUM	numeral	55
PRON	pronoun	294
PROPN	proper noun	1,755
SCONJ	subord. conjunction	24
VERB	verb	567,512
X	epenthetic consonants	3

Table 1: Category distribution in Loflòc

Form	Lemma	PoS
seguda	seguda	NOUN
seguda	segut	ADJ
seguda	segudar	VERB
seguda	sèire	VERB
seguda	sèser	VERB
seguda	sègre	VERB

Table 2: Loflòc entries for *seguda*

4. General Coverage Analysis

Corpus	Dialect	# Tokens	Coverage (%)	# Types	Coverage (%)
Restaure	Gascon	3,311	67.02	1,367	51.06
	Lengadocian	3,608	91.57	1,424	83.92
	Provençau	1,085	85.90	544	77.94
	Lemosin	1,975	76.20	941	60.47
	All	9,979	79.77	3,636	62.95
Tolosa TB	Gascon	3,465	69.38	1,351	51.89
	Lengadocian	16,192	91.06	4,314	79.37
	Provençau	1,113	87.51	539	79.04
	Lemosin	1,147	74.54	559	60.47
	All	21,917	86.59	5,941	69.77
OcWikiAnnot	All	1,812,127	82.25	143,160	30.56

Table 3: Coverage of existing annotated corpora

Corpus	Dialect	# Tokens	rightPOS (%)	wrongPOS (%)
Restaure	Gascon	3,311	56.45	10.57
	Lengadocian	3,608	87.83	3.74
	Provençau	1,085	81.75	4.15
	Lemosin	1,975	72.51	3.70
	All	9,979	73.72	6.04
Tolosa TB	Gascon	3,465	60.49	8.89
	Lengadocian	16,192	89.22	1.84
	Provençau	1,113	83.92	3.59
	Lemosin	1,147	72.10	2.44
	All	21,917	83.52	3.08

Table 4: Token-level coverage of annotated corpora taking into account the POS annotation

5. Perspectives

Our three priorities for the future of Loflòc are as follows:

- Adding **detailed morphological information** to the current version of the lexicon
- Extending the **coverage of other dialects** by relying on existing lexicographic resources
- Including **other spelling norms**, starting with the Mistralian norm, which is widely used in Provençau

Acknowledgements: We would like to thank Benaset Dazeas and Aure Séguier of Lo Congrès Permanent de la Lenga Occitana, who helped us in the first steps of Loflòc construction. This work has been carried out within the framework of several projects : the ANR-14-CE24-0003 RESTAURE project and the ANR-21-CE27-0004 DIVITAL project supported by the French National Research Agency; the EFA 227/16 LINGUATEC Project, financed by the POCTEFA Interreg European funds. The work of Aleksandra Miletić was funded by the Academy of Finland through project No. 342859 “CorCoDial – Corpus-based computational dialectology”